The Quantitative Evaluation of Functional Neuroimaging Experiments

(Bias-Variance Tradeoffs in the NPAIRS Data Analysis Framework)

Stephen C. Strother[1,2,3,6], Lars Kai Hansen[4], Jon R. Anderson[1,3], Stephen LaConte[6], Ulrik Kjems[4], Rafal Kustra[5], John Sidtis[2], Sally Frutiger[2], Suraj A. Muley[2], Essa Yacob[1,7], Xiaoping Hu[1,7], David Rottenberg[1,2,3]

[1]Radiology and [2]Neurology Departments., University of Minnesota
[3]PET Imaging Center, VA Medical Center, Minneapolis
[4]Department of Mathematical Modeling, Technical University of Denmark
[5]Health Sciences, University of Toronto and Statistics, Stanford University
[6]Biomedical Engineering, University of Minnesota
[7]Center for Magnetic Resonance Research, University of Minnesota

Principal Funding Sources: NIH Human Brain Project, P20-MH57180 and Danish Research Councils for the Natural and Technical Sciences.

The Problem

- We collect a set of scans (high-dimensional multivariate image vectors) with an unknown spatio-temporal structure.
- Each scan is acquired under one of a finite set of experimental design conditions or brain states.
- PROBLEM: How to determine the spatio-temporal structure that "best" describes the variation among these experimental brain states?

The Philosophy

"I believe in ignorance based methods because humans have a lot of ignorance and we should play to our strong suit."

Eric Lander
Whitehead Institute, M.I.T.

The field of statistical learning or machine learning provides a coherent scientific approach for this viewpoint!
See: Machine Learning for Science: State of the art and future prospects. E. Mjolsness & D. DeCoste. Science, 293:2051-2055, September, 2001

Presentation Overview

- **AIMS.** *Optimize functional neuroimaging results by measuring* internal result consistency *and* bias-variance tradeoffs *without relying on:*
  - < Prior neuroscientific expectations, i.e., the neuroscientific-bias problem;
  - < Inferential tests based on Maximum Likelihood (ML) parameter estimates.
- **METHODS.** *NPAIRS Quality Metrics for Multiple PET Tasks:*
  - < Prediction error in a flexible, multivariate, cross-validation framework;
  - < Statistical parametric map (SPM) reproducibility;
  - < Provide a data-driven alternative to ROC curves.
- **RESULTS.** *BOLD-fMRI, within-subject, run-to-run comparisons as a function of preprocessing:*
  - < Exploratory canonical variates analysis (CVA) to establish signal subspace;
  - < Prediction vs. reproducibility plots as f(model complexity);
  - < Preprocessing optimization for group vs. individual subjects.
- **CONCLUSIONS.**

Why Test Internal Result Consistency?
(The Neuroscientific Bias Problem)

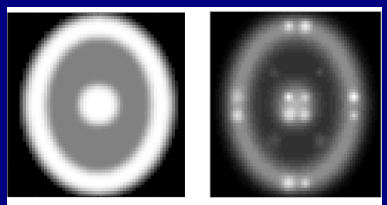
- Patterns of functional activation obtained from neuro-imaging studies reflect interactions between choices of:
  - < Research question and activation task;
  - < Experimental design parameters;
  - < A complex series of data-analysis-chain choices including:
    - – Data acquisition and post-acquisition preprocessing;
    - – Data-analysis model selection and tunning model complexity.
- Generation of a plausible result is often taken as justification for the choices made, leading to a *systematic bias in the field towards prevailing neuroscientific expectations*.

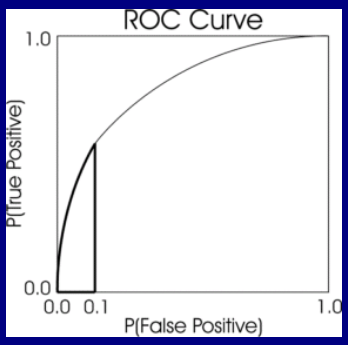
Are Maximum Likelihood Parameter Estimates Enough?

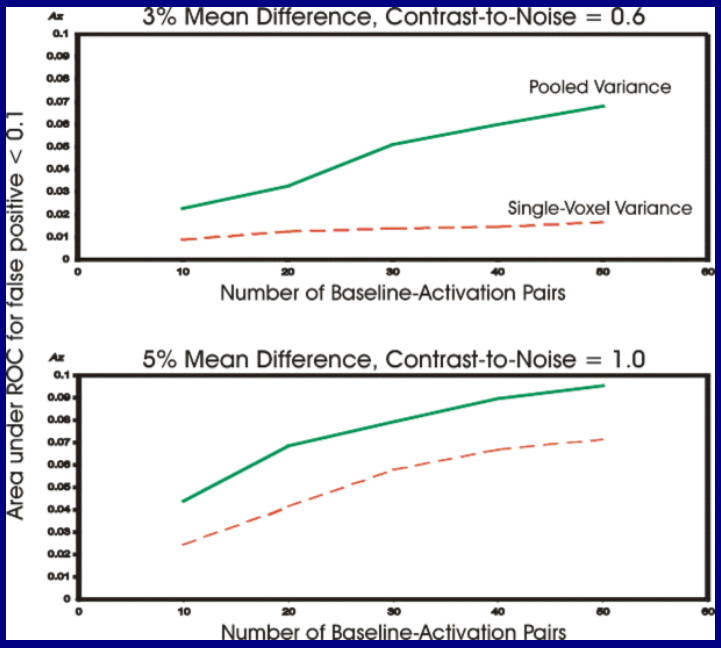
- Perhaps asymptotically, but:
  - < ML focuses on asymptotically unbiased, minimum-variance estimates;
  - < This ignores the bias-variance tradeoffs inherent in parameter estimates from finite samples;
  - < We have no idea what large . asymptotic means in functional neuroimaging;
  - < There are better signal detectors than ML estimates that are asymptotically-biased but have smaller parameter variance;
  - < **In real, finite data sets there is a bias-variance tradeoff to be considered, even for the t-test!**

Is the *t*-test efficient?

ROC Simulations







(See: Lukic AS, Wernick MN, Strother SC. An evaluation of methods for detecting brain activations from PET or fMRI images. A.I. Medicine (Invited paper, in press))

The NPAIRS Framework

- Uses split-half resampling to provide:
  - < measurements of prediction (generalization) error and SPM pattern reproducibility (reliability);
  - < uncorrelated signal and noise SPMs from any data analysis model;
  - < a reproducible SPM (rSPM) on a common statistical Z-score scale;
  - < implicitly includes random observation effects, e.g., subjects, runs;
  - < a measure of individual observation influence.

< **N** Nonparametric
< **P** Prediction
< **A** Activation
< **I** Influence
< **R** Reproducibility
< **S** reSampling

The NPAIRS Framework

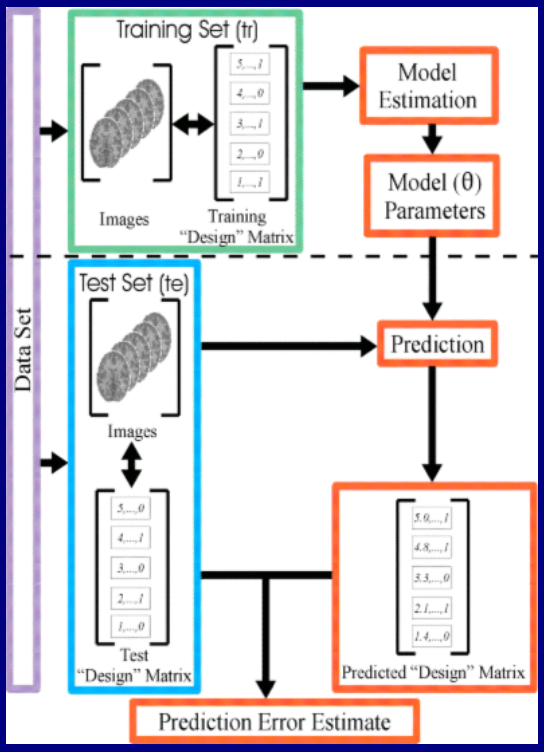
- The development and application of the NPAIRS framework is described in the following papers:

< Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Siditis J, Frutiger S, Muley S, LaConte S, Rottenberg D. The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *Neuroimage* (in press).

< Kjems U, Hansen LK, Anderson J, Frutiger SA, Sidtis JJ, Rottenberg D, Strother SC. The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. *Neuroimage* (in press).

< LaConte S, Anderson J, Muley S, Frutiger S, Hansen LK, Yacoub E, Xiaoping H, Rottenberg D, Ashe J, Strother SC. Evaluating preprocessing choices in single-subject BOLD-fMRI studies using data-driven performance metrics. *Neuroimage* (submitted).

NPAIRS: Prediction/Crossvalidation reSampling



Prediction Metrics in Functional Neuro-imaging Studies.
Morch N, Hansen LK, Strother SC, Svarer C, Rottenberg DA, Lautrup B, Savoy R, Paulson OB. Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. In: Duncan J, Gindi G, eds: Lecture Notes in Computer Science 1230: Information Processing in Medical Imaging. Springer-Verlag, 1997, 259-270.

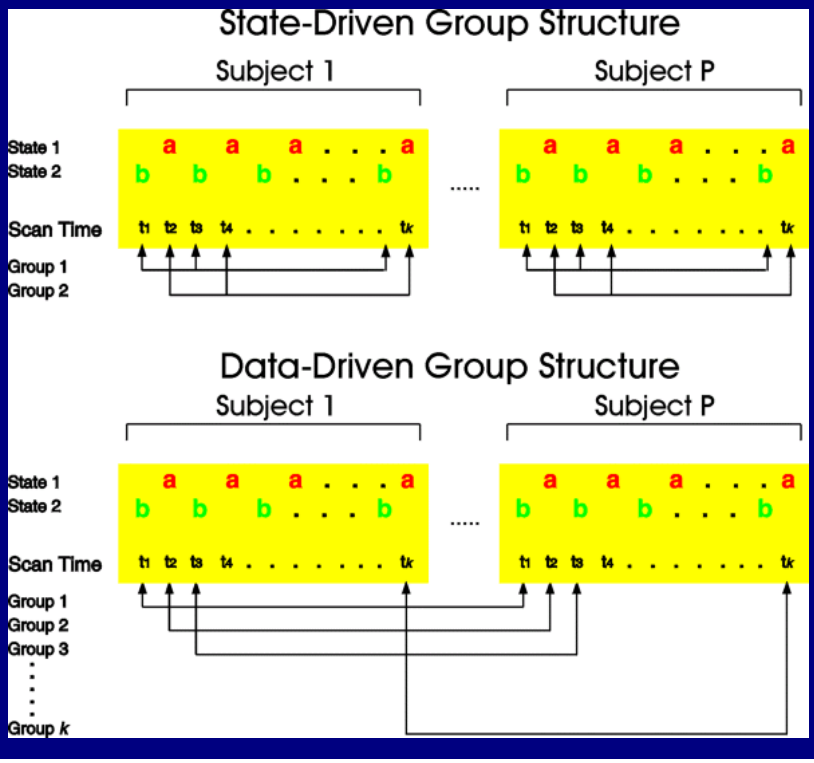
Hansen LK, Larsen J, Nielsen FA, Strother SC, Rostrup E, Savoy R, Lange N, Sidtis J, Svarer C, Paulson OB. Generalizable patterns in neuroimaging: How many principal components?. *Neuroimage,* 9:534-544, 1999.

Kustra R, Strother SC. Penalized discriminant analysis of [^{15}O]water PET brain images with prediction error selection of smoothing and regularization hyperparameters. *IEEE Trans Med Img* 20:376-387, 2001.

A Flexible Multivariate Framework for NPAIRS

- Canonical Variates Analysis (Mardia et al., 1979):
  - < Closely related to Linear Discriminant Analysis, Canonical Correlation Analysis and Partial Least Squares;
  - < Maximises the multivariate signal-to-noise ratio of (Between-Group)/(Pooled Within-Group) covariance;
  - < Provides approximate correction for random subject effects;
  - < Efficiently detects mean AND spatial interaction signals;
  - < Easily vary model complexity.
    - – utilizing experimental state-driven or data-driven group structures;
    - – preprocessing with different types/numbers of basis functions.

Experimental- vs. Data-Driven Analysis

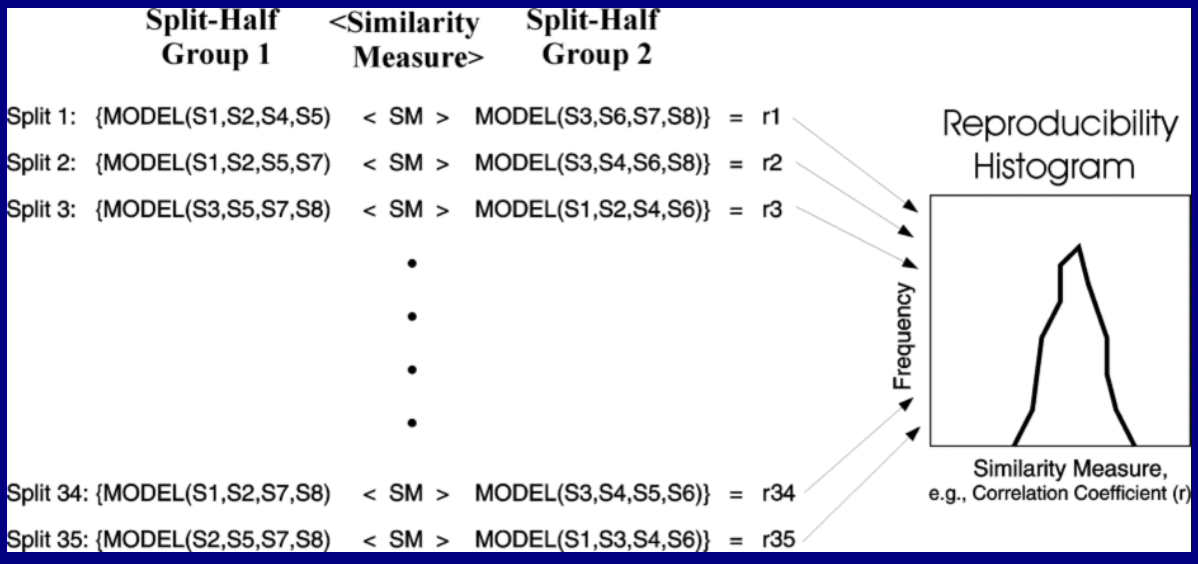


Predicting the Brain State with CVA

$$p(g^{(j)} \mid \mathbf{x}_{te}^{(j)}; \theta_{tr}) = \frac{1}{C} \exp\left\{ -\frac{1}{2} \left\| \mathbf{L}_{tr}^{T} \left(\mathbf{U}_{tr}^{*} \right)^{T} \left(\mathbf{x}_{te}^{(j)} - \bar{\mathbf{x}}_{tr}^{(g^{(\bullet)})} \right) \right\|^{2} \right\} p(g^{(j)})$$

- Identifies the regions needed to explain systematic variations between scans by linearly combining with a new scan to predict the experimental state of the brain, i.e. the group of the new test scan.
- The probability of predicting the group, g, of a new scan $p(g^{(j)} \mid \mathbf{x}_{te}^{(j)}; \theta_{tr})$
- Is a weighted, multivariate Gaussian distribution $\frac{1}{C} \exp\left\{ -\frac{1}{2} \| \ \|^{2} \right\} p(g^{(j)})$
- Dependent on the Euclidean distance $\| \ \|^{2}$
  - < Between the training group mean and the new scan $\left(\mathbf{x}_{te}^{(j)} - \bar{\mathbf{x}}_{tr}^{(g^{(\bullet)})} \right)$
  - < Projected onto a set of non-orthogonal canonical eigenimages $\mathbf{L}_{tr}^{T} \left(\mathbf{U}_{tr}^{*} \right)^{T}$
  - < With flexibly choosen type and number of basis functions $\mathbf{U}_{tr}^{*}$

NPAIRS: Split-half reSampling to Obtain Activation-Pattern Reproducibility Metrics



For 8 subjects {S1, ..., S8} each of 35 splits creates two groups of 4 subjects from which any model produces two independent SPMs for comparison.

NPAIRS: Reproducibility Metrics in Functional Neuroimaging Studies

Strother SC, Lange N, Anderson JR, Schaper KA, Rehm K, Hansen LK, Rottenberg DA. Activation pattern reproducibility: Measuring the effects of group size and data analysis models. Hum Brain Mapp, 5:312-316, 1997.

Strother SC, Rehm K, Lange N, Anderson JR, Schaper KA, Hansen LK, Rottenberg DA. Measuring activation pattern reproducibility using resampling techniques. In: Quantitative functional brain imaging with Positron Emission Tomography. (Carson RE, Daube-Witherspoon ME, Herscovitch P, eds.), Academic Press, San Diego, pp. 241-246, 1998.

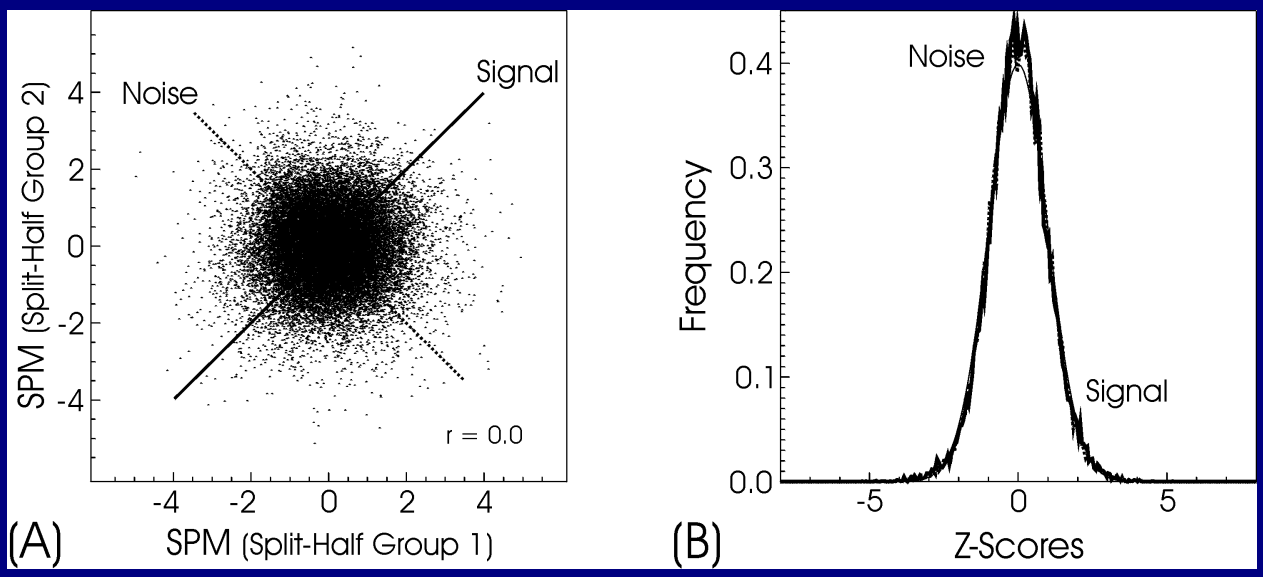
Tegeler C, Strother SC, Anderson JR, Kim S-G. Reproducibility of BOLD-based functional MRI obtained at 4T. Hum Brain Mapp, 7:267-283, 1999.

Frutiger S, Strother SC, Anderson JR, Sidtis JJ, Arnold JB, Rottenberg DA. Multivariate predictive relationship between kinematic and functional activation patterns in a PET study of visuomotor learning. Neuroimage 12:515-527, 2000.

Muley SA, Strother SC, Ashe J, Frutiger SA, Anderson JR, Sidtis JJ, Rottenberg DA. Effects of changes in experimental design on PET studies of isometric force. Neuroimage 13:185-195, 2001.

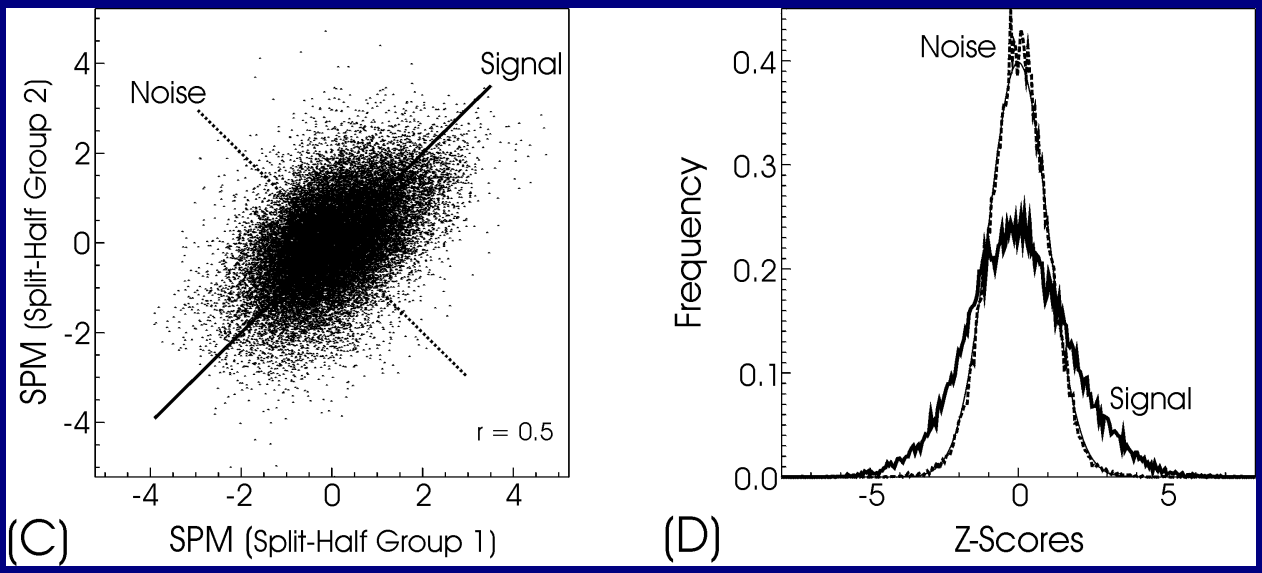
Shaw M, Strother SC, McFarlane AC, Morris P, Anderson J, Clark CR, Egan GF. Abnormal functional connectivity in post-traumatic stress disorder. Neuroimage (in press).

NPAIRS: Z-Scored, Reproducing SPM (rSPM(Z)) for Independent, NonReproducing SPMs



PCA of scatter-plot correlation matrix: $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$

NPAIRS: Z-Scored, Reproducing SPM (rSPM(Z)) for Independent, Reproducing SPMs

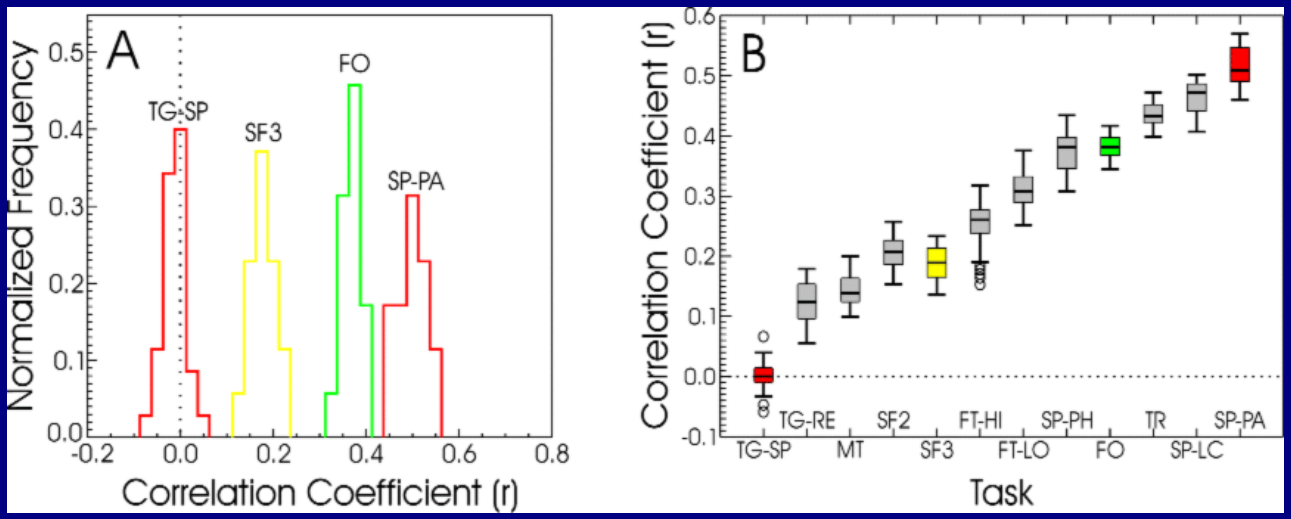


PCA of scatter-plot correlation matrix: $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$

Twelve PET Data Sets

- **12 PET data-analysis sets (8-subjects/set) performing 11 tasks:**
  - *3 Speech tasks:*
    - Syllable production, **SP-PA**
    - Lip closure, **SP-LC**
    - Sustained phonation, **SP-PH**
  - *2 Figure tracing tasks:*
    - Standard tracing, **TR**
    - Standard tracing followed by mirrored tracing, **MT**
  - *3 Finger movement tasks with auditory pacing:*
    - Finger opposition at 1 Hz, **FO**
    - Finger tapping, low amplitude, parametrically varied 0-3 Hz, **FT-LO**
    - Finger tapping, high amplitude, parametrically varied 0-3 Hz, **FT-HI**
  - *2 Parametricaly varied static force tasks with visual feedback:*
    - Alternating control-force design, **SF2**
    - Randomized block design, **SF3**
  - *1 Target interception (circular moving target within an annular path):*
    - Contrast reaction type (button push vs. joystick) independent of speed, **TG-RE**
    - Contrast speed (fast vs. slow) independent of reaction type, **TG-SP**

NPAIRS: Reproducibility Histograms for Twelve PET Tasks



(See: Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Siditis J, Frutiger S, Muley S, LaConte S, Rottenberg D. The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *Neuroimage* (in press))

Within-Subject fMRI Comparisons

(LaConte et al., Neuroimage, 2002 (submitted))

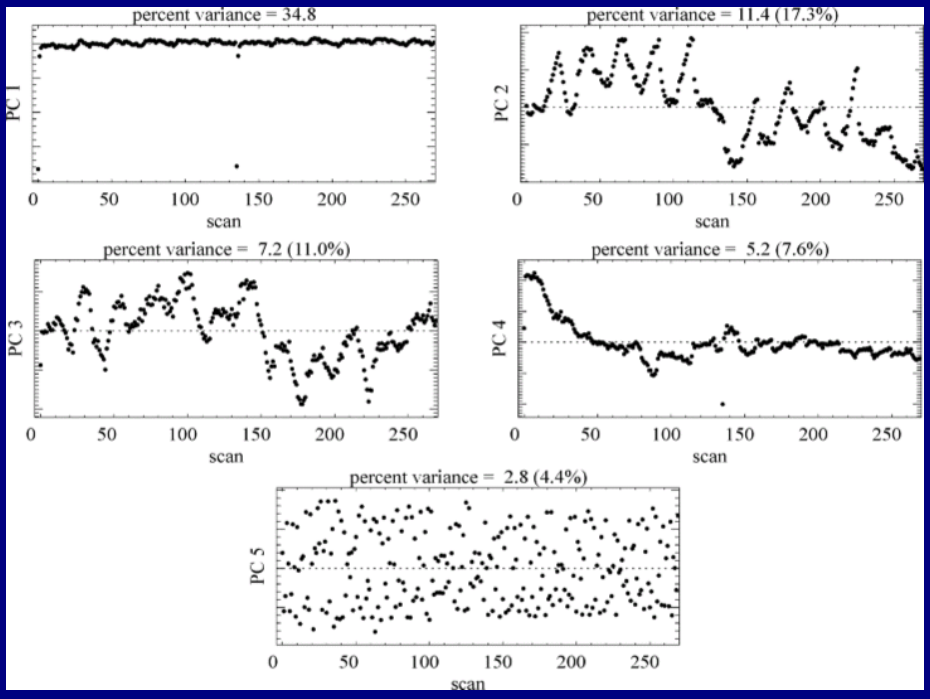
< Sixteen subjects with 2 runs/subject

< Acquisition:
- Whole-brain, 1.5T BOLD-EPI;
- 30 slices = 1 whole-brain scan;
- 1 oblique slice = 3.44 x 3.44 x 5 mm^3;
- TR/TE = 4000 ms/70 ms

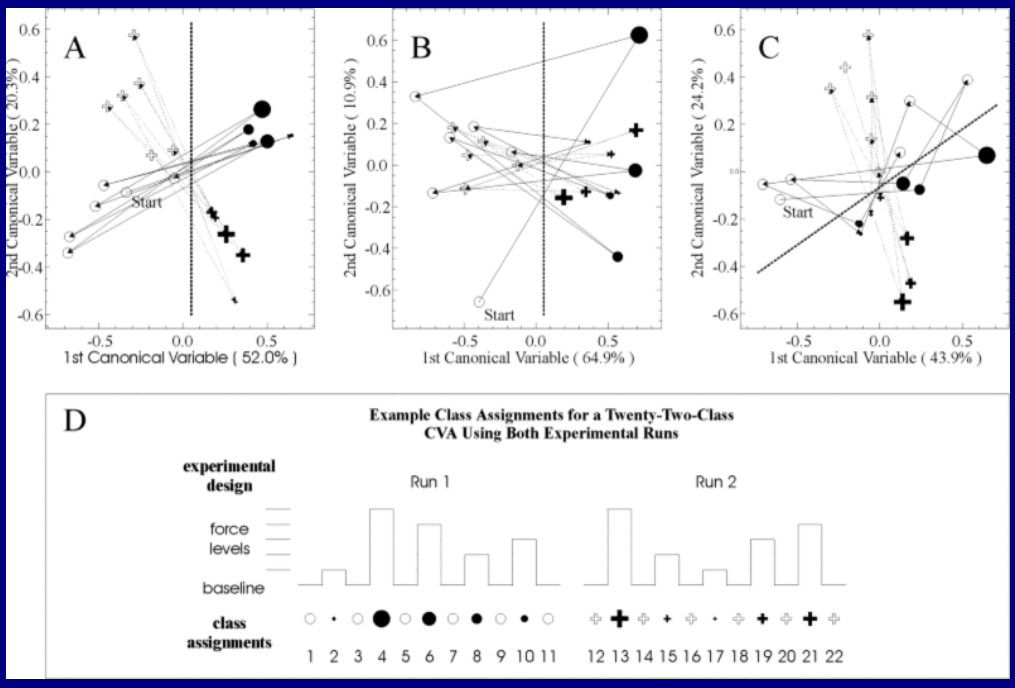
< Experimental Design:
- Parametric static isometric force (sf);
- Run: [5 x (b_1, ... , b_{11}, $sf\#_1$, ... , $sf\#_{11}$), b_1, ... , b_{11}] = 121 scans;
  - sf1=200g, sf2=400g, sf3=600g, sf4=800g, sf5=1000g.

< Analyzed with PCA and Penalized CVA (Kustra & Strother, IEEE TMI, 20:376-387, 2001):
- 22-group and 2-group analyses;
- Dropped initial non-equilibrium and state-transition scans.

PCA Preprocessing for Two-Run Static Force



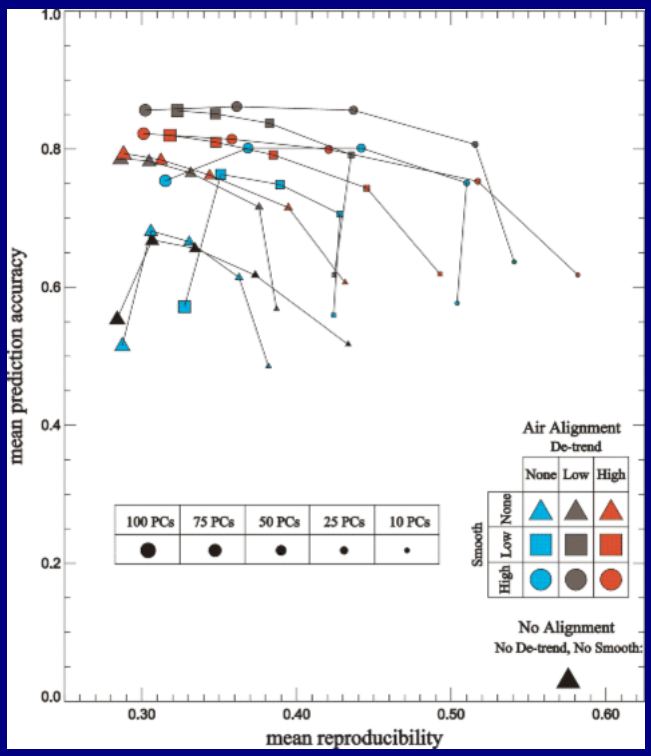
fMRI Static Force: Exploratory CVA for Three Subjects



Preprocessing for Two-Run Static Force

- ***All runs/subject(s) passed initial quality control:***
  - movement (AIR 3) < 1 voxel;
  - no artifacts in functional or structural scans;
  - no obvious outliers in PCA of centered data matrix.
- ***Within-Subject Alignment:***
  - None;
  - Across runs using AIR 3.08 to 1st scan of run one.
- ***Temporal Detrending using GLM Cosine Basis (a la SPM):***
  - DT0: None;
  - DTL: 0.5-cosine/run;
  - DTH: 0.5, 1.0, 1.5 and 2.0 cosines/run.
- ***Spatial Smoothing with 2D Gaussian:***
  - GS0: None;
  - GSL: FWHM = 1.5 voxels = 0.52 mm;
  - GSH: FWHM = 6 voxels = 21 mm.

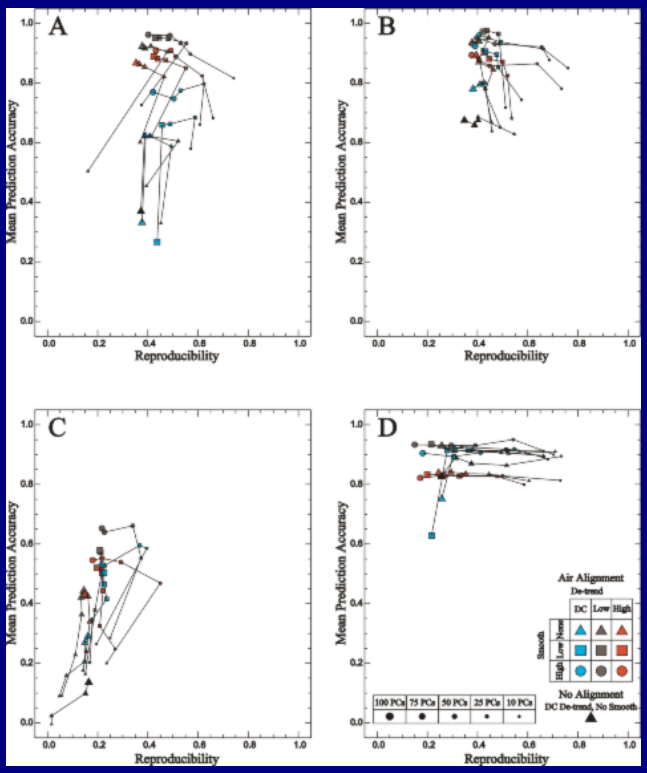
Multi-Subject Prediction Accuracy vs Reproducibility



1. A Classic Bias-Variance Tradeoff. As model complexity increases (#PCs 10 6 100) prediction of design matrix's class labels improves and reproducibility (i.e., activation SNR) decreases.

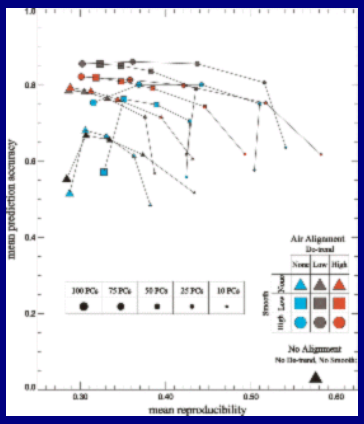
2. Optimizing Performance. Like an ROC plot there is a single point, (1, 1), on this prediction vs. reproducibility plot with the best performance; at this location the model has perfectly predicted the design matrix while extracting an 4 SNR.

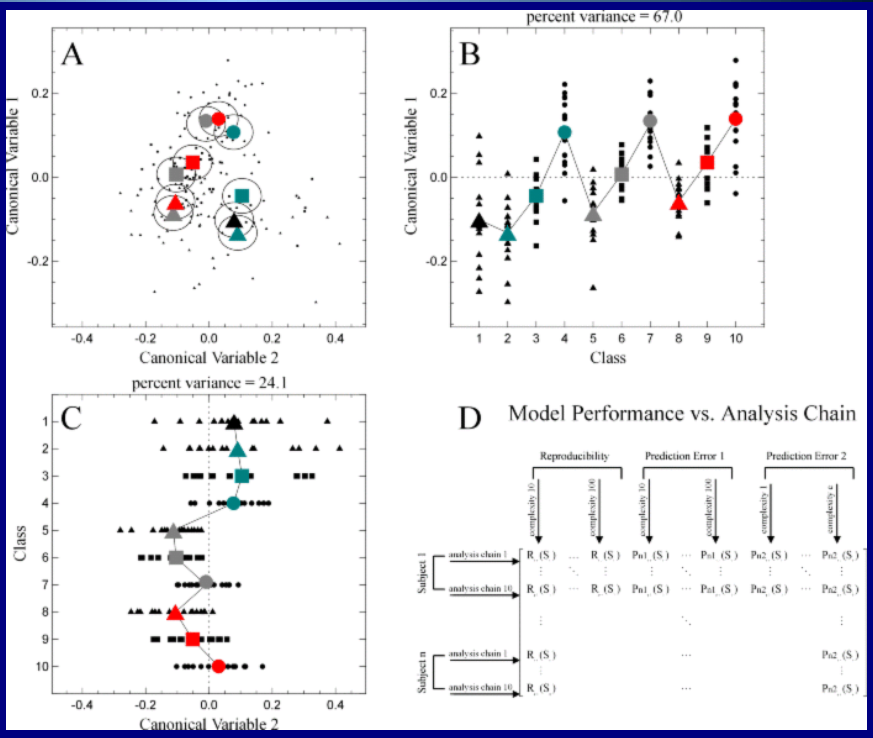
Single-Subject Prediction Accuracy vs. Reproducibility



CVA of Model Performance vs. Analysis Chain

Mean Prediction
vs.
Reproducibility





The NPAIRS Framework

- Provides a framework for determining the spatio-temporal structure that "best" describes the variation among experimental brain states.
- Provides a non-parametric approach for maximizing the SNR of spatial activation patterns (i.e., their reproducibility) while allowing for random effects and controlling for model generalization ability.
- Provides a data-driven alternative to "true simulations" for ROC curves.
- May be readily applied within and between subjects, laboratories, modalities and tasks.
- Benefits are:
  - < "semi-orthogonal" scales for quantitatively ranking experimental and methodological choices within and between different classes of models.
  - < experimental and methodological optimization is no longer strongly dependent on particular data analytic model assumptions.
  - < replaces *result validity based only on* inferential p-values and neuroscientific expectations.